On estimation of Polish real estate market characteristics using Internet data sources

Maciej Beręsewicz (maciej.beresewicz@ue.poznan.pl) 1,2

Keywords: representativeness, real estate market, web scraping, Internet data sources.

1. INTRODUCTION

New data sources became an important issue in the Official Statistics as well as in statistics in general. This topic was raised in the context of usage of register data for deriving statistics but also as a source of auxiliary variables for small area estimation. However, assumed full coverage of population by registers/administrative records do not indicate that it reflects the current state nor fully covers the information needs of society.

Recently with rapid growth of new technologies (e.g. mobiles phones) or Internet coverage it becomes clear that new potential data source for statistics is rising [1]. Term big data appeared in the context of massive data sets that were time consuming to analyse with existing technology. Big data also describes the process how data is generated and this issue is more important for statisticians than what is the volume of the data sets.

Initial classification of big data was proposed in the UNECE project big data for Official Statistics [2] that includes human-sourced information, process-mediated data and machine-generated data (internet of things). Nonetheless there is a specific group of sources that could be classified as a Internet data sources which are defined [3] as data collected and maintained by units external to statistical offices and administrative regulations, and are (mainly) available on the Internet (through web-based databases).

From the methodological point there are questions that are crucial in the context of statistical data source – what is potential usefulness of use of this data? Is the data representative for general population? In what percent data source cover the general population? What statistics could be derived from the new data sources? What is the quality of data and furthermore what methods of estimation could be applied to receive sufficient statistics?

For example, Polish real estate market is only partially covered by official statistics that are published by National Bank of Poland with co-operation of Central Statistical Office. Reports are presented with delay (e.g. statistics on 2013 were published in the beginning of the October 2014) and in limited scope. On the other hand brokers and owners of properties in order to sell or rent need to publish information on the Internet to get to the potential buyers or reinters. Potential usefulness of the Internet data sources should be assessed.

The aim of the paper is to present data sources on real estate market in Poland and assesses the possibility of use of new data sources for deriving statistics on secondary real estate market. The problem of estimation of population size (secondary real estate market) will be presented and approach proposed by [4] will be applied.

1 Department of Statistics, Poznan University of Economics

2 Center for Small Area Estimation, Statistical Office in Poznan
2. PROPOSED APPROACH

For the purpose of the study web-scraping technique will be applied to obtain the data from the selected online portals on real estate market in Poland. Special program written in R using XML, RCurl and httr packages was developed and data for Poznań, Poland was collected. Data cleaning part included text processing and record linkage to impute and identified duplicated entries.

To estimate size of secondary real estate market in Poznań, Poland Dual System Estimator (DSE) will be discussed. Capture-Recapture estimators or DSE are widely used in the official statistics in order to estimate size of the population or evaluate the coverage of census/registers [4, 5]. In the context of many data sources that are on the Internet the problem of estimation of population size also appears. Nonetheless, there are several issues that differs the new data sources from the existing statistical sources (e.g. census, registers) which one of them is coverage, in particular undercoverage and overcoverage errors. Recently [4] proposed an approach that includes in the dual system estimator coverage error with respect to registers and census data.

In order to estimate the size of secondary real estate market in Poznan approach proposed by [4] will be adopted. The approach applies log-linear model that takes into account overcoverage in the first source and assumes that second is free of coverage errors. For the purpose of the study similar situation will be assumed. Simulation study will be presented to assess the variance of proposed [4] estimator in the context of Internet data sources and secondary real estate market in Poznan, Poland. Results of the study could be further studied for real estate market research.

3. FINAL REMARKS

Internet data sources and big data were noticed by statisticians and official statistics in the context of statistical data source. Despite many advantages connected to low level of aggregation and timeliness the data sources should be studied in detail in order to assess the quality and potential usefulness for statistics. In the paper approach using the Internet data sources for estimation size of Polish secondary real estate market will be presented and discussed.

REFERENCES


